Robust Estimation of Distance Between Sets of Points

Rodrigo Moreno, Sandeep Koppal and Ebo de Muinck

Linköping University Post Print

N.B.: When citing this work, cite the original article.

Original Publication:

Rodrigo Moreno, Sandeep Koppal and Ebo de Muinck, Robust Estimation of Distance Between Sets of Points, 2013, Pattern Recognition Letters.

<http://dx.doi.org/10.1016/j.patrec.2013.08.012>

Copyright: Elsevier

<http://www.elsevier.com/>

Postprint available at: Linköping University Electronic Press

<http://urn.kb.se/resolve?urn=urn:nbn:se:liu:diva-97282>

Robust Estimation of Distance Between Sets of Points

Rodrigo Moreno^{a,b,*}, Sandeep Koppal^{a,b}, Ebo de Muinck^{a,b}

^a*Center for Medical Image Science and Visualization (CMIV), Linköping University*

^b*Department of Medical and Health Sciences (IMH), Linköping University*

Abstract

This paper proposes a new methodology for computing Hausdorff distances between sets of points in a robust way. In a first step, robust nearest neighbor distance distributions between the two sets of points are obtained by considering reliability measures in the computations through a Monte Carlo scheme. In a second step, the computed distributions are operated using random variables algebra in order to obtain probability distributions of the average, minimum or maximum distances. In the last step, different statistics are computed from these distributions. A statistical test of significance, the nearest neighbor index, in addition to the newly proposed divergence and clustering indices are used to compare the computed measurements with respect to values obtained by chance. Results on synthetic and real data show that the proposed method is more robust than the standard Hausdorff distance. In addition, unlike previously proposed methods based on thresholding, it is appropriate for problems that can be modeled through point processes.

Keywords: Spatial statistics, distance estimation, Hausdorff distance, nearest neighbor distance distribution

1. Introduction

Distance is arguably the most important parameter in spatial statistics. While distance between two points can be defined as the shortest path between them with respect to a metric, there is not a unique way for defining distance between sets of points. Actually, several distance measurements have been proposed in the bibliography (see below). In general, advantages and drawbacks of a distance measurement have to be considered in order to assess its appropriateness for a specific application.

Classical distance measurements between sets of points include the minimum and maximum distances between all possible pairs of points. However, these measurements are sensitive to noise. Although robust alternatives are the distance between the two centroids and the Mahalanobis' generalized distance (Everitt et al., 2011), they can yield results that disagree with intuition. For example, this distance is close to zero for points located at two concentric rings, disregarding the distances between the rings. A widely used alternative is the Hausdorff distance (HD). However, the main drawback of this measurement is its lack of robustness, which makes it inappropriate for noisy input data. In this context, this paper is focused on applications where robustness of the HD is an issue.

Some efforts have been done in order to tackle this robustness problem. For example, Huttenlocher et al. (1993) use robust statistics, and Knauer et al. (2011) consider imprecisions in the input data to compute more robust measurements of HD. Unfortunately, these methods are also affected by noise (cf. Section 3).

Indeed, assuming that a measure of reliability of every single point at the input data is available, that is the probability of its occurrence, such a measure can be used to improve the robustness of distance measurements. Following this approach, different fuzzy extensions of the HD have been proposed (Chaudhuri

*Corresponding author: Tel.: +46 10-10 39579

Email addresses: rodrigo.moreno@liu.se (Rodrigo Moreno), sandeep.koppal@liu.se (Sandeep Koppal), ebo.de.muinck@liu.se (Ebo de Muinck)

and Rosenfeld, 1999; Fan, 1998; Boxer, 1997). These extensions have in common that they model reliability maps as fuzzy sets. Their strategy is to compute distance as a combination of different HDs computed between sets of points generated through thresholding-like strategies from these fuzzy maps. Although modeling reliability through membership degrees can be appropriate for some applications, this is not the case for point processes, where this concept is related to probability of occurrence of events rather than to membership degrees.

This problem can be better explained through an example. Lets us consider two locations p_1 and p_2 of the same set of points with the same measure of reliability. Thresholding on the reliability map can be interpreted as the conditional probability of occurrence of an event at p_1 given that an event happened at p_2 is 100% and the same conditional probability is 0% provided that an event at p_2 does not happen. In other words, thresholding imposes a perfect correlation between events p_1 and p_2 , which is inappropriate if every event is considered independent, or at least if they are not perfectly correlated.

An example where the perfect correlation is not a good assumption is given by microscopic images of tissues stained through immunohistochemistry. For these images, the presence of stains are usually associated with the detection of target proteins at a specific location and not with its local concentration as a threshold-based method would assume. As it will be shown, although threshold-based methods can yield robust measurements, the extra assumption of perfect correlation they make can lead to large overestimations of the distance measurements. This makes threshold-based methods inappropriate for applications such as the analysis of immunohistochemistry images.

Although the methods proposed by Zhao et al. (2005) and Vivek and Sudha (2006, 2007) do not use fuzzy sets theory, they also follow a thresholding-like approach and consequently share the same aforementioned limitations of the methods based on fuzzy logic.

In this context, this paper proposes a robust variant of the HD where reliability is modeled as the probability of occurrence of events. Our strategy is based on robust estimations of probability distributions of distance measures which take into account reliability maps of the input data. These probability distributions are modeled as random variables. In a final step, robust statistics are used to estimate statistical parameters of the computed probability distributions. The main advantage of the proposed method compared to previous robust approaches is that its scope of application is broader provided that it does not use any a priori correlation assumption between events, making it appropriate for both point processes and problems where reliability can be modeled through membership degrees.

The paper is organized as follows. Section 2 describes the basics of the Hausdorff distance. Section 3 presents the proposed methodology to estimate distances. Section 4 discusses statistical significance tests for the proposed measurements. Section 5 describes considerations regarding implementation. Section 6 shows experiments in simulated and real data. Finally, Section 7 discusses the results and makes some final remarks.

2. The Hausdorff Distance

This section introduces the basic concepts related to the Hausdorff distance, which will be used in the proposed method presented in the next section.

Let A and B be two sets of points of cardinality $card(A)$ and $card(B)$ respectively, and $d(x, y)$ be the distance between points x and y with respect to a specific metric. The HD, D_H , can be computed as:

$$D_H(A, B) = \max\{h(A, B), h(B, A)\}. \quad (1)$$

with $h(A, B)$ being the directed Hausdorff distance (DHD), which in turn is given by:

$$h(A, B) = \sup_{x \in A} \inf_{y \in B} d(x, y). \quad (2)$$

Notice that, in order to become a pseudo-metric, D_H combines independently computed distances both from points in A to points in B and from points in B to points in A , which are generally different. Nevertheless,

Figure 1: Two sets of points (marked with circles and squares) each of them with a noisy measurement in the middle of the other set. Although, in average, the two sets of points are far away from each other, the HD is small due to these two noisy measurements.

in some applications only directed distances are of interest, so in those cases the combination of distances performed in (1) can be omitted.

For convenience, HD can be written in terms of the nearest neighbor distance distribution (NNDD). The empirical cumulative distribution function (CDF) of the NNDD between A and B , F , can be estimated as (Diggle, 2003; Stoyan, 2006):

$$F(t; A, B) = \frac{1}{\text{card}(A)} \sum_{x \in A} H \left(t - \inf_{y \in B} d(x, y) \right), \text{ for } t \geq 0, \quad (3)$$

where H is the Heaviside step function. Let $F^{-1}(z; A, B) = \inf_{t \in \mathbb{R}} \{F(t; A, B) \geq z\}$ be the generalized inverse of a CDF. Consequently, (2) can be rewritten as:

$$h(A, B) = \min\{t \mid F(t; A, B) = 1\} = F^{-1}(1; A, B). \quad (4)$$

As shown in equations (1) to (4), h , and consequently D_H , can be biased by noise, due to the use of non-robust statistics such as the minimum, maximum, infimum and supremum. Thus, D_H is only appropriate for noiseless input data. A natural approach for tackling this drawback is to use robust estimators on the NNDD instead. Thus, for example, the following extension of the DHD was proposed by Huttenlocher et al. (1993):

$$h_{k/q}(A, B) = k^{th} \inf_{y \in B} d(x, y) \quad (5)$$

where k^{th} stands for the k -th q -quantile of F . Equivalently, this equation can be written in terms of the NNDD as:

$$h_{k/q}(A, B) = \min\{t \mid F(t; A, B) \geq k/q\} = F^{-1}(k/q; A, B). \quad (6)$$

As an example, $h_{1/2}$ corresponds to the median of F .

It is clear from (6) that the use of quantiles helps but does not completely solve the robustness problem, since $h_{k/q}$ still use the non-robust infimum statistic required in (3). This issue can be seen in the example of Figure 1. Although the two depicted sets of points are far away from each other, the estimated HD is counterintuitively small due to a couple of noisy points, for both h and $h_{k/q}$.

Unfortunately, replacing the infimum in (3) by robust statistics is not a good option since the estimated distance is required to be related to minimum distances. Thus, even using the first decile instead of the infimum can lead to overestimations of the distance. Instead, we propose to tackle the problem by using reliability maps as described in the next section.

3. Computing Hausdorff distances with reliability measures

Robustness of the estimation of the NNDD can be improved by considering reliability maps of the measures in the formulation. These reliability maps can be seen as local probability measurements of non-

noisiness and are assumed to be given (for example, this would be the case when they can directly be obtained from the acquisition system) or estimated in a preprocessing step.

Let r_A and r_B be reliability maps of the input sets of points. Thus, the reliability maps can be included in the computations in order to obtain robust estimations of the NNDD. The following formula is proposed for this purpose:

$$F_r(t; A, B) = \frac{\sum_{x \in A} r_A(x) P(\hat{d}(x, B) \leq t)}{\sum_{x \in A} r_A(x)}, \text{ for } t \geq 0, \quad (7)$$

where $\hat{d}(x, B)$ is a random variable that models the variability due to the reliability map r_B of the infimum distance between x and the points in B , and P is the probability operator. Notice that if $r_B = 1$ for all points in B , $\hat{d}(x, B)$ becomes equivalent to the infimum distance between x and the points in B and the probability operator becomes the Heaviside step function of (3). However, due to r_B , such a distance is not unique. Instead, it has a particular probability distribution in a certain range of distance values (actually, such a range is defined by the minimum and maximum distance from x to any point at B) that can be modeled through a random variable for every x . In turn, as seen in the equation, r_A is used for computing a weighted mean of $P(\hat{d}(x, B) \leq t)$. The idea with this is to give more importance in the global distance computations to the most reliable points in A .

Notice that a closed formulation of $\hat{d}(x, B)$ is difficult to obtain. Despite this, good estimations can be obtained by running Monte Carlo simulations where points y in B are filtered out in every run by comparing $r_B(y)$ with randomly generated numbers before computing distances between points. Finally, robust extensions of the DHD and its improved version proposed by Huttenlocher et al. (1993) can be obtained by replacing F by F_r in (4) and (6) respectively. Implementation details of the proposed method are given in Section 5.

The final step is to compute the non-directed HD from DHDs. It is important to take into account for this task that estimated statistical parameters convey less information than complete probability distributions, since parameters reduce the informative content of a distribution to a single value. Thus, it is advantageous to perform such estimations as the very last step. In the case of computing HD, this means that it is convenient to delay the k -th q -quantile estimation to the last step. Such a delay can be performed as follows.

First, assume independence between $F_r(t; A, B)$ and $F_r(t; B, A)$. Since both NNDDs $F_r(t; A, B)$ and $F_r(t; B, A)$ can be seen as random variables, it is possible to generate a new random variable for the maximum between both distributions. It can be shown that the distribution of such random variable is given by (Papoulis and Pillai, 2002):

$$F_{max}(t; A, B) = F_r(t; A, B) F_r(t; B, A). \quad (8)$$

Finally, a robust HD can be obtained by computing the k -th q -quantile of F_{max} . A further advantage of this approach is that it is easy to estimate additional statistical parameters. For example, the median absolute deviation (MAD) of F_{max} can be used as a measure of variability of the distance estimations, which is straightforward to compute from the distribution F_{max} .

Notice that the proposed methodology is general in the sense that $F_r(t; A, B)$ and $F_r(t; B, A)$ can be combined in any other way before the estimation of statistical parameters. For example, computing random variables for the minimum and the average can be used to compute the range of values of distance. The distribution of the minimum can be computed as (Papoulis and Pillai, 2002):

$$F_{min}(t; A, B) = F_r(t; A, B) + F_r(t; B, A) - F_r(t; A, B) F_r(t; B, A), \quad (9)$$

In turn, the probability density function (PDF) of the average distribution, f_{avg} , can be computed as:

$$f_{avg}(t; A, B) = (f_r(A, B) * f_r(B, A))(2t) \quad (10)$$

where f_r is the probability density function of F_r and “*” is the convolution between both distributions (Grinstead and Snell, 1997). In line with the method proposed by Huttenlocher et al. (1993), the proposed method leaves k and q as parameters that should be selected depending on the application.

4. Statistical Analysis

Distance measurements must be complemented with statistical significance tests. In the case of the DHD, these tests compare the computed distances between A and B to the distances between A and a set of randomly located points C with similar characteristics as B . In practice, this means that $\text{card}(B) = \text{card}(C)$ and $r_B(y_i) = r_C(z_i)$ for $y_i \in B$ and $z_i \in C$. The NNDDs between A and C can be estimated in a similar way as described above.

If the median is used to compute DHDs, the Mann-Whitney U-test can be applied for checking the statistical significance of the difference between both distances: A to B and A to C . Alternatively, if quantiles are used, more general tests can be applied, such as the one proposed by Ozturk and Balakrishnan (2009).

An additional standard way to compare distance measurements between the two sets of points and distances that would be expected on the basis of chance is the nearest neighbor index (NNI) (Clark and Evans, 1954; Ripley, 2004; Cressie, 1993). Given a distance measurement D , the NNI can be computed as:

$$\text{NNI} = \frac{D(A, B)}{E\{D(A, B)\}}, \quad (11)$$

where the denominator is the expected value of the distance between A and B . It is difficult to obtain a closed formulation for computing this expected value, especially for geodesic distances, since A and B can have different characteristics. Alternatively, this value can be estimated by computing the distance between two sets of randomly located points C_A and C_B , generated with similar features as A and B respectively. Intuitively, NNI takes values smaller or larger than one when A and B appear more clustered or more disperse than expected by chance respectively.

Furthermore, we propose the following directional extensions of the NNI:

$$\text{SI} = \frac{D(A, B)}{E\{D(A, B)|A\}}, \quad (12)$$

$$\text{GI} = \frac{D(A, B)}{E\{D(A, B)|B\}}, \quad (13)$$

where the denominators are the expected values of the distance between A and B when A and B are fixed respectively. Following the methodology described above, these expected values can be estimated by computing the distance from A to C_B , and from C_A to B respectively. Intuitively, SI, which stands for spatial divergence index, is related to the spatial divergence between A and B measured from the perspective of A . Thus, spatially correlated and divergent sets of points will have values of SI smaller and larger than one respectively. In turn, GI, which stands for grouping index, can be seen as a index of clustering of A from the perspective of B . GI is usually larger or equal to one, and it increases when the points of A appear in clusters.

5. Implementation Issues

The efficiency of the method relies on efficient computations of NNDDs between sets of points, since this process is run T times, where T is the number of Monte Carlo simulations. Indeed, computing point-to-point distances in order to estimate the NNDDs is only a possibility when cardinalities of both A and B are small. Thus, a more efficient approach is to use distance transforms (DT). Thus, for computing the DHD from A to B the DT is initialized with the selected points from A , that is, the distance is set to zero at those points, and the method for estimating the DT is run until all points at B are covered by the transform. In

the case of Euclidean distances, there are efficient methods for computing the DT in linear time (Meijster et al., 2002). For geodesic distances, it is possible to compute DT through fast marching schemes (Sethian, 1999; Osher and Fedkiw, 2003) using the same aforementioned initialization. The PDF of the NNDD of the i -th Monte Carlo simulation can be approximated through a weighted histogram of the distances at A with r_A being the weighting function. The final estimated PDF is the mean of the NNDD computed at every simulation. Algorithm 1 shows the pseudocode of the proposed method for estimating the empirical PDF of the NNDD between two sets of points. Notice that, although it is possible to computing $\hat{d}(x, B)$ pointwise, this is not needed to estimate the NNDD, saving in that way computational resources.

Algorithm 1 Estimation of the empirical PDF of the NNDD between A and B

```

1: function NNDD( $A, B, r_A, r_B$ )
2:   Initialize  $f_r$ 
3:   for  $i = 1 \rightarrow T$  do ▷  $T$  is the number of iterations
4:     Generate a random number  $M_B(y) \in [0, 1]$  for every  $y \in B$ 
5:     Select  $B' = \{x \in B | r_B(y) \geq M_B(y)\}$ 
6:     Compute the distance transform  $DT$  from  $B'$ 
7:     Compute a weighted histogram  $l_A$  with  $DT(x \in A)$  and weights  $r_A$ 
8:      $f_r \leftarrow f_r + l_A$ 
9:   end for
10:  Normalize  $f_r$ 
11:  return  $f_r$ 
12: end function

```

It is important to remark that the same DT in every simulation can be used for estimating both NNDDs: from A to B and from A to C , by measuring the distances at different positions. An additional advantage of using DTs is that it allows the method to be generalized to other type of input data such as lines or closed regions, since the initialization step of the DT is independent of the type of the input data. Regarding the operations on random variables, computing the maximum, minimum or average distribution is computationally inexpensive.

6. Experimental Results

The following subsections describe the experiments conducted in order to test the proposed measurements.

6.1. Experiment with Synthetic Data

In the first experiment, two clusters of points with different densities have randomly been sampled in order to generate the first set of points A , and the second set of points B has been generated by mirroring A as shown in Figure 2. The distances between the internal clusters and between the external ones are 150 and 750 respectively. The external clusters have 120 points while the internal ones have 480. Different levels of noise have been added in order to evaluate the robustness of the measurements. Noisy levels correspond to the number of added noisy points given as a fraction of the original number of points in the dataset. Local density of points have been used as reliability maps. In addition to the HD, the methods proposed by Huttenlocher et al. (1993), Chaudhuri and Rosenfeld (1999) and Vivek and Sudha (2006) referred to as HDH, HDF and HDG respectively have been coded for comparison. T has been set to 500 provided that the variability of the results for the proposed method is negligible after this number of Monte Carlo simulations.

Figure 3 shows the evolution of different distance measurements with the level of noise. As expected, the estimations through HD and HDH are not robust, as their values are largely reduced even with low noise levels. In turn, although HDF and HDG yield robust measurements, these largely overestimate the expected distance measurement due to the fact that these methods largely penalize areas with a density of points below the highest density in the space. Actually, these measurements are more related to the distance

Figure 2: Synthetic dataset. Every set of points is composed by a small and a big cluster.

Figure 3: Performance of distance measurements with respect to the amount of added noisy points.

between the external clusters, since the applied thresholding tends to exclude the internal clusters in the computations, disregarding the fact that the internal clusters have much more points than the external ones. In practice, the large bias introduced by these two methods make them inappropriate for problems that cannot accurately be modeled through membership functions. As shown in the figure, the proposed measurements have a better performance than the other tested approaches. The median of F_{max} and F_{min} , denoted by $\hat{F}_{max}$ and $\hat{F}_{min}$, can be used as the expected range of values of the distance, while the median of F_{avg} , denoted as $\hat{F}_{avg}$, can be used to get robust estimations of the distance between the two sets of points.

6.2. Experiments with Real Data

The following example shows how the proposed method can be used to analyze stained slides taken from atherosclerotic plaques extracted from patients through carotid endarterectomy. The slides are stained through immunohistochemistry with specific antibodies to target the hypoxia-inducible factor 1 α (HIF1 α) and Prolyl Hydroxylase 3 (PHD3) proteins. The spatial relationship between these two proteins is important to understand the development of atherosclerotic plaques in humans. Figure 4 shows a preprocessed slide imaged through a fluorescence microscope with an isotropic resolution of 1.02 μm . As in the previous example, local density of points have been used as reliability maps, 500 Monte Carlo iterations have been run and distance transforms have been computed through fast marching in order to consider geodesic distances inside the tissue. This makes the distance measurements to consider the effect of boundaries and/or holes in the tissue.

From visual assessment, one can expect the distance from HIF-1 α (A) to PHD3 (B) to be smaller than from B to A , since B is more homogeneously distributed throughout the tissue than A . Moreover, one can expect NNIs greater than one due to the clusters of both stains. In addition, values of SI for the distance from A to B are expected to be smaller than one, since all clusters of B are relatively close to a cluster in A . Notice that the opposite is not always true, for instance, some of the clusters of A around the cavity (the lumen) are not close to any cluster at B . Furthermore, B is expected to have a larger value of GI, since it has fewer clusters than A . Finally, distances beyond 3 mm do not correspond with the intuition given by visual assessment.

Figure 4: Stained slide of an atherosclerotic plaque. Blue indicates presence of HIF1 α , while red represents presence of PHD3. The boundaries of the tissue are marked in black. The slide has been preprocessed for a better visualization.

Table 1: Directional distance estimations, NNI, SI and GI for the slide of Figure 4. A and B correspond to HIF1 α and PHD3 respectively. $\hat{F}_r$ is the median of the robust estimation of directional NNDD. Distances are given in microns.

Method	A to B	NNI	SI	GI	B to A	NNI	SI	GI
HD	1,079.20	3.19	0.72	3.47	1,143.14	1.87	1.05	1.63
HDH	137.19	1.37	0.44	1.40	157.19	1.33	0.61	1.36
HDF	3,696.25	9.75	0.47	16.61	11,802.18	26.03	0.80	48.49
HDG	4,449.61	44.53	0.78	41.01	10,548.80	89.43	1.29	73.89
$\hat{F}_r$	229.59	1.27	0.33	1.15	882.33	4.22	0.97	3.68

Table 1 show the directional distances estimated through different methods with their corresponding NNI. First, HD and HDH yield similar distances from B to A and from A to B , while the robust variants show a large difference between them, which is more consistent with intuition. In addition, the short distances reported by HDH also seem unlikely. Moreover, similarly to the synthetic dataset, HDF and HDG tend to largely overestimate the distance estimations. These overestimations have also a large impact in the reported NNI and GI values. Regarding NNIs, HD and HDH yield counterintuitive results as the NNI for the distance from A to B should be smaller than the NNI for the distance from B to A . This observation is also valid for the GI. Finally, all values of SI are consistent with respect to visual assessment. In summary, the proposed method yields results that better correspond with visual assessments.

Regarding statistical significance, the Mann Withney U-test run on the proposed measurements indicates that both directional distances are statistically different from distances computed from a type of stain to a randomly generated stain ($p < 0.001$). It is important to remark that, unlike the proposed method, no statistical tests have been proposed for HDF and HDG hitherto. Finally, Figure 5 shows the NNDDs computed through the proposed method for the stained image. After operating these NNDDs, the estimated values of $\hat{F}_{min}$, $\hat{F}_{avg}$ and $\hat{F}_{max}$ are 178.59, 614.57 and 984.33 μm respectively. It is noteworthy to remark that, although the median (i.e., $k = 1$ and $q = 2$), was appropriate for estimating statistics from the PDFs for the tested slide, other quantiles could be more appropriate in other applications.

As already mentioned, the main advantage of the proposed method is that, unlike previously proposed extensions of the Hausdorff distance such as HDH, HDG or HDF, it is appropriate for problems that can be modeled as point processes, such as the ones shown in the previous experiments. Certainly, the proposed method is also applicable to other types of problems, for example, to those where reliability is related to membership levels as seen in the following example. The task is to determine the level of similarity/dissimilarity between every couple of images, which must be related to distance measurements estimated between them. To this task, gray-scale edge maps were generated from the input images through

Figure 5: NNDDs computed through the proposed method. (a) HIF1 α to PHD3. (b) A random stain with similar characteristics to PHD3. (c) PHD3 to HIF1 α . (d) A random stain with similar characteristics to PHD3 to a random stain with similar characteristics to HIF1 α .

the Canny edge detector without the non-maximum suppression and hysteresis thresholding steps with a standard deviation of one (Canny, 1986). Notice that a gray-scale edge map can be interpreted as a membership function of every pixel to the edge category. HDF, HDG and $\hat{F}_{avg}$ have been applied to every pair of gray-scale edge maps in order to estimate distances between the images.

Table 2 shows the images used in the experiment, which were taken from the ORL database of faces with mild illumination changes (Samaria and Harter, 1994). Table 2 also shows the distances computed through the tested methods. As shown in the table, the proposed method correctly classify all pairs of images as similar or dissimilar by applying a simple threshold of 5 pixels, with the exception of the distance between I_4 and I_6 . In turn, selecting a threshold for performing such a classification from the distances computed through HDF or HDG is very difficult, which can also be seen in the table.

6.3. Computational Complexity

Regarding computational complexity, let $O(K)$ be the complexity of computing the distance transform. First, the complexity of HD and HDH is exactly $O(K)$. Second, the complexity of HDF is $O(MK)$ where M is the number of different fuzzy membership values considered in the computations. Third, the complexity of HDG is $O(QK)$ where Q is the number of gray levels after quantization. Finally, the complexity of the proposed method is $O(TK)$, with T being the number of Monte Carlo iterations..

7. Discussion

This paper proposed a new methodology for computing Hausdorff distances between sets of points in a robust way. Reliability measures were considered to compute robust directed NNDDs through Monte Carlo simulations. In a second step, the computed directed NNDD are operated using random variables algebra in order to obtain distributions of the average, minimum or maximum distances. In the last step, quantiles are used to estimate parameters from these distributions. Statistical tests of significance, the NNI and the newly proposed SI and GI were used to compare the distance estimations with respect to values obtained by chance. Results show that the proposed method is more robust than the standard HD and HDH. In addition, its results are not biased when the perfect correlation between events required by threshold-based methods, such as HDF and HDG, is not complied. This broadens the scope of use of the new method beyond the problems where reliability can be modeled through membership degrees.

It is important to mention that the proposed method can straightforwardly be extended for computing local measurements of distance by performing the computations with respect to a sliding window. In this case, edge effects have to be considered in the computations, since it is well-known that the use of windows can lead to overestimation of the distances (Stoyan, 2006; Cressie, 1993). As an example, in Stoyan (2006) the Hanisch’s estimator (Hanisch, 1984) is recommended for performing such a correction.

Furthermore, similarly to previous approaches, the proposed measurements are not metrics. However, the field of application of these measurements is still huge as many applications do not require distance estimations to be metrics. For example, this restriction is not relevant for assessing spatial relationships between stained cells. Ongoing research includes testing the proposed measurements in different applications,

Table 2: Distance between edge maps extracted from the shown images estimated through HDF, HDG and $\hat{F}_{avg}$. Distances are given in pixels.

		I ₂	I ₃	I ₄	I ₅	I ₆
HDF	I ₁	67.11	14.40	61.33	37.11	44.52
	I ₂		72.94	73.30	57.70	74.12
	I ₃			59.65	37.01	35.24
	I ₄				54.00	61.42
	I ₅					27.73
HDG	I ₁	76.58	28.16	70.40	70.86	66.12
	I ₂		79.97	82.76	96.55	87.09
	I ₃			66.53	85.73	71.34
	I ₄				73.04	66.94
	I ₅					62.88
$\hat{F}_{avg}$	I ₁	2.88	3.38	6.38	8.88	7.88
	I ₂		4.88	6.38	9.38	8.13
	I ₃			5.88	8.13	8.38
	I ₄				5.13	4.88
	I ₅					5.13

especially in biology. The proposed method allows for the statistical analysis of spatial relationships between different cell types. This will generate novel insights into how these cells interact.

8. Acknowledgments

This research has been partially supported by the Hjärt Lungfonden, grant No. 20100460.

References

- Boxer, L., 1997. On Hausdorff-like metrics for fuzzy sets. *Pattern Recognition Letters* 18, 115–118.
- Canny, J., 1986. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 8, 679–698.
- Chaudhuri, B.B., Rosenfeld, A., 1999. A modified Hausdorff distance between fuzzy sets. *Information Sciences* 118, 159–171.
- Clark, P.J., Evans, F., 1954. Distance to nearest neighbor as a measure of spatial relationships in populations. *Ecology* 35, 445–453.
- Cressie, N.A.C., 1993. *Statistics for Spatial Data*. Wiley-Interscience.
- Diggle, P.J., 2003. *Statistical Analysis of Spatial Point Patterns*. Hodder Education Publishers. 2nd edition.
- Everitt, B.S., Landau, S., Leese, M., Stahl, D., 2011. *Cluster Analysis*. Wiley. 5th edition.
- Fan, J., 1998. Note on Hausdorff-like metrics for fuzzy sets. *Pattern Recognition Letters* 19, 793–796.
- Grinstead, C.M., Snell, J.L., 1997. *Introduction to Probability*. AMS. 2nd edition.
- Hanisch, K.H., 1984. Some remarks on estimators of the distribution function of nearest-neighbour distance in stationary spatial point patterns. *Statistics* 15, 4009–412.
- Huttenlocher, D.P., Klanderman, G.A., Rucklidge, W.J., 1993. Comparing images using the Hausdorff distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 15, 850–863.
- Knauer, C., Löffler, M., Scherfenberg, M., Wolle, T., 2011. The directed Hausdorff distance between imprecise point sets. *Theoretical Computer Science* 412, 4173–4186.
- Meijster, A., Roerdink, J.B., Hesselink, W., 2002. A general algorithm for computing distance transforms in linear time, in: Goutsias, J., Vincent, L., Bloomberg, D.S. (Eds.), *Mathematical Morphology and its applications to image and signal processing*. Kluwer, pp. 331–340.
- Osher, S.J., Fedkiw, R.P., 2003. *Level Set Methods and Dynamic Implicit Surfaces*. Springer.

- Ozturk, O., Balakrishnan, N., 2009. Exact two-sample nonparametric test for quantile difference between two populations based on ranked set samples. *Annals of the Institute of Statistical Mathematics* 61, 235–249.
- Papoulis, A., Pillai, S.U., 2002. *Probability, Random Variables and Stochastic Processes*. McGraw-Hill. 4th edition.
- Ripley, B.D., 2004. *Spatial Statistics*. Wiley-Interscience.
- Samaria, F., Harter, A., 1994. Parameterisation of a stochastic model for human face identification, in: *Proc. IEEE Workshop on Applications of Computer Vision*.
- Sethian, J., 1999. *Level Set Methods and Fast Marching Methods Evolving Interfaces in Computational Geometry, Fluid Mechanics, Computer Vision, and Materials Science*. Cambridge University Press. 2nd edition.
- Stoyan, D., 2006. On estimators of the nearest neighbour distance distribution function for stationary point processes. *Metrika* 64, 139–150.
- Vivek, E.P., Sudha, N., 2006. Gray Hausdorff distance measure for comparing face images. *IEEE Transactions on Information Forensics and Security* 1, 342–349.
- Vivek, E.P., Sudha, N., 2007. Robust Hausdorff distance measure for face recognition. *Pattern Recognition* 40, 431–442.
- Zhao, C., Shi, W., Deng, Y., 2005. A new Hausdorff distance for image matching. *Pattern Recognition Letters* 26, 581–586.